USO DE FERRAMENTAS DE *DATA MINING* NA IDENTIFICAÇÃO DO NÍVEL DE ESTABILIDADE DE TALUDES EM ATERRO

USE OF DATA MINING TOOLS FOR STABILITY CONDITION IDENTIFICATION OF SOIL EMBANKMENTS

Tinoco, Joaquim; *ISISE, Universidade do Minho, Guimarães, Portugal,* jtinoco@civil.uminho.pt
Gomes Correia, António; *ISISE, Universidade do Minho, Guimarães, Portugal,* agc@civil.uminho.pt
Cortez, Paulo; Algoritmi, *Universidade do Minho, Guimarães, Portugal,* pcortez@dsi.uminho.pt

RESUMO

Para uma eficiente gestão das infraestruturas viárias (rodo-ferroviárias), nomeadamente da rede de taludes, é essencial proceder ao levantamento do nível de estabilidade dos diferentes elementos, nomeadamente dos taludes. Este levantamento, de caracter periódico, permitirá uma quantificação e priorização dos recursos disponíveis para ações de manutenção e conservação da rede. Por outro lado, a informação recolhida necessária para a categorização dos taludes deverá permitir uma classificação realista do nível de estabilidade do mesmo e ao menor custo possível. Esta tarefa tem-se revelado complexa e até ao momento apenas parcialmente concluída. Usualmente procura-se ponderar o nível de fiabilidade da classificação atribuída a cada talude e o investimento realizado na recolha da informação necessária para a classificação. Neste trabalho é apresentada uma proposta de avaliação do nível de estabilidade de taludes em aterro. O sistema proposto utiliza informação recolhida durante inspeções de rotina, por norma de fácil obtenção, bem como todo um conjunto de características geométricas e geográficas do talude e atribui uma classe ao nível de estabilidade do talude em análise. Esta proposta, desenvolvida através da aplicação de ferramentas de *data mining*, procura maximizar a informação disponível visando uma classificação realista do nível de estabilidade do talude. Os resultados obtidos utilizando um conjunto significativo de taludes da rede ferroviária britânica evidenciam um bom desempenho dos modelos propostos na identificação do nível de estabilidade de um determinado talude com base em informação essencialmente visual. De sublinhar ainda o contributo dos modelos propostos para uma gestão mais eficiente dos recursos disponíveis ao nível das ações de manutenção e conservação da rede de taludes.

ABSTRACT

Keeping in mind an efficient management of the transportation infrastructures, namely the slopes network, it is fundamental to measure the stability condition of the different element of the network. This information will allow a more efficient management of the available budgets for maintenance tasks of the network. Moreover, it is important that the information used to define the stability condition of each slope allow a reliable classification and at the same time be easy to collect and cheap as possible. This task has proved to be complex and so for only partially concluded. Until now what have been done is to try to balance the reliability of the defined stability condition and the investment needed to get the information required by the classification systems. In this work a new approach to measure the stability condition of soil embankments is presented. The proposed system is feed with information that can be easily obtained through visual routine inspections, as well as geographic and geologic data, and calculate the stability condition level of a given soil embankment. This proposal was developed through the application of data mining tools and aims to maximize all available information in order to get the stability condition of a soil embankment as realistic as possible. Based on a representative database from the railway network of the UK, the achieved results shows a good performance of the proposed models in stability condition identification of a given soil embankment, using as model attributes almost only visual information. It is also important to underline the contribution of the proposed models for a more efficient management of the available budgets for maintenance and repair tasks.

1 - INTRODUÇÃO

Após um longo período de investimento e desenvolvimento, Portugal dispõe atualmente de uma rede de infraestruturas de transporte, nomeadamente rodo e ferroviária, bastante completa. O desafio atual prende-se com a manutenção da rede existente de forma a assegurar todas as condições de segurança e mobilidade. Face ao elevado número de elementos constituintes da rede e das limitações orçamentais disponíveis para gestão de toda a infraestrutura, torna-se fundamental dispor de um conjunto de ferramentas que auxiliem os gestores responsáveis nas suas tarefas de forma a otimizar os recursos disponíveis.

Um dos elementos que requerem particular atenção, necessitando de uma observação/manutenção com regularidade, é a rede de taludes que constitui a rede rodo e ferroviária. A falta de manutenção pode levar à ocorrência de deslizamentos/derrocadas com graves perdas económicas e humanas. Por outro lado, a gestão de toda a rede representa um custo significativo para as respetivas concessionárias. Torna-se portanto fundamental desenvolver um conjunto de ferramentas capazes de identificar o nível de estabilidade de um determinado talude preferencialmente através de informação recolhida durante as inspeções de rotina. Desta forma será possível priorizar intervenções, minimizando os custos de manutenção e a ocorrência de acidentes.

Embora existam alguns sistemas para a previsão da ocorrência de deslizamentos/derrocadas, uma parte significativa tem como alvo os taludes naturais, não sendo adequados à avaliação de taludes feitos pelo homem. Além disso, estes sistemas apresentam como principal desvantagem o facto de requerem informação por vezes de difícil obtenção, como por exemplo através de ensaios específicos ou equipamentos de monitorização dispendiosos, acrescentando ainda o facto de em alguns casos terem sido desenvolvidos utilizando informação proveniente de casos de estudo muito concretos, limitando um pouco o respetivo domínio de aplicação. Há ainda a sublinhar o facto de alguns sistemas serem caracterizados por uma forte componente de subjetividade. A título de referência, é de sublinhar os sistemas propostos por Cheng e Hoang (2014) onde é apresentado um método de classificação que calcula a probabilidade de rotura de um determinado talude. Também Ahangar-Asr et al. (2010) propuseram um modelo para a determinação do fator de segurança de taludes em rocha e em solo, através da aplicação de técnicas de *data mining* (DM). Lu e Rosenbaum (2003) e Sakellariou e Ferentinou (2005) fizeram também uso de técnicas de DM mas neste caso para prever diretamente se um determinado talude iria ruir ou não. Todas estas três abordagens, embora tenham conseguido um desempenho satisfatório, apresentam como principal limitação o facto de terem de assumir *a priori* o tipo de rotura do talude. Além disso, foram desenvolvidos utilizando uma base de dados com um número de registos bastante reduzido. Mais recentemente, um novo sistema foi apresentado por Pinheiro et al. (2015). O sistema SQI (*Slope Quality Index*), caracterizado pela sua grande flexibilidade, assenta na avaliação de diferentes fatores que influenciam o comportamento de um talude, sendo aplicável quer a taludes em rocha quer em solo. Através da ponderação dos diferentes fatores obtém-se uma classificação final do talude, representativa do nível de estabilidade do mesmo.

O comportamento de um talude depende de um elevado número de fatores, alguns deles de difícil avaliação (AGC, 2007), (Fay et al., 2012). Por outro lado, existem atualmente poderosas ferramentas capazes de explorar grandes volumes de dados e extrair conhecimento útil. Estas ferramentas, usualmente conhecidas por DM, têm sido aplicadas com sucesso em diversas áreas do conhecimento, nomeadamente na área de Engenharia Civil e em particular em geotecnia (Tinoco et al., 2014a), (Miranda et al., 2011). Ao nível do estudo da estabilidade de taludes, Gavin e Xue (2009) calcularam o índice de fiabilidade de um determinado talude, bem como a posição do nível freático com recurso a algoritmos genéticos. Também Wang et al. (2005) avaliaram a estabilidade de um talude através da aplicação de redes neuronais artificiais. Duas outras propostas desenvolvidas através da aplicação de máquinas de vetor de suporte foram apresentadas por Cheng et al. (2012) e Yao et al. (2008).

O presente trabalho, tem como principal objetivo o uso de ferramentas de inteligência artificial no desenvolvimento de um sistema de identificação do nível de estabilidade de taludes em aterro, alimentado por informação recolhida durante inspeções de rotina (informação visual) e complementada com alguma informação geométrica, geológica e geográfica. Para questões de treino/validação do sistema foi utilizada uma base de dados relativa à rede de taludes da rede ferroviária do Reino Unido, disponibilizada pela NetworkRail. O problema em estudo foi abordado seguindo duas estratégias distintas: como um problema de classificação nominal; convertido num problema de regressão. Para cada uma destes estratégias foram aplicados dois algoritmos de DM, nomeadamente as Redes Neuronais Artificiais (RNAs) e as Máquinas de Vetores de Suporte (MVSs).

2 - BASE DE DADOS

Um elemento fundamental para a realização de qualquer estudo assente na aplicação de ferramentas de DM é a existência de uma base de dados representativa do problema em estudo. A proposta apresentada neste trabalho para a identificação do nível de estabilidade de taludes em aterro foi desenvolvido com uma base de dados constituída por 25673 registos, tendo sido disponibilizada pela NetworkRail e é relativa à rede ferroviária de Inglaterra.

O modelo apresentado para a classificação do nível de estabilidade de um determinado talude em aterro, daqui em diante designado por EHC (“Earthwork Hazard Categorization”) é alimentado por 53 variáveis usualmente medidas durante inspeções de rotina. Abaixo são apresentadas algumas das variáveis consideradas:

- Altura;
- Inclinação;
- Geologia da base;
- Proteção da superfície;
- Drenagem subterrânea;
- Drenagem dos terrenos envolventes;
- Atividade animal;
- Existência de construções na base;
- Existência de árvores;
- etc.

Adicionalmente às variáveis acima mencionadas e às restantes utilizadas no desenvolvimento da proposta aqui apresentada, é sabido que muitas outras têm uma forte influência no estudo da estabilidade de taludes, como por exemplo a quantidade de precipitação durante um determinado período de tempo. No entanto, como tal informação não está disponível na base de dados utilizada no presente estudo, as mesmas não foram consideradas. De sublinhar contudo, que um dos objetivos deste estudo passa por tentar desenvolver um sistema de identificação do nível de estabilidade de taludes em aterro utilizando essencialmente informação visual recolhida durante inspeções de rotina e que seja de fácil obtenção. Além disso, a proposta apresentada tem como principal objetivo dar apoio à tomada de decisão do ponto de vista de gestão de uma rede de taludes, não se pretendendo uma análise detalhada da estabilidade de um determinado talude.

A cada registo da base de dados está atribuída uma classe EHC, constituída por 4 níveis (A, B, C e D), onde A representa um elevado nível de estabilidade e D corresponde a um nível de estabilidade com uma probabilidade de rotura superior. A definição da classe atribuída a cada talude resulta da experiencia dos Engenheiros e Técnicos da NetworkRail e será assumida como representativa do real nível de estabilidade do talude.

A Figura 1 ilustra a distribuição dos 25673 registos da base de dados pelas 4 classes EHC. Da sua análise é possível observar uma assimétrica bastante pronunciada, a qual terá um efeito preponderante na resposta dos modelos para cada uma das classes, tal como analisado e discutido mais detalhadamente na secção 4. Embora do ponto de visto de aprendizagem dos modelos tal distribuição assimétrica da informação tenha um efeito negativo, a mesma é representativa da realidade, tendo em conta que é espectável que uma parte significativa dos taludes da rede apresente um nível de estabilidade elevado (classe A), e que apenas alguns apresentem um elevada probabilidade de rotura (classe D).

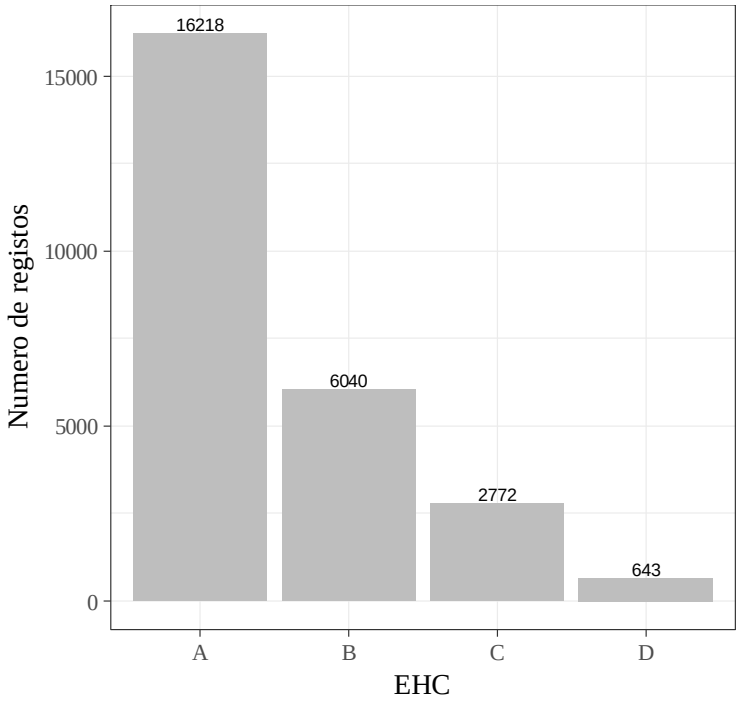


Figura 1 - Distribuição do número de registos pelas 4 classes EHC

3 - METODOLOGIA E ALGORITMOS

Como referido anteriormente, o sistema proposto para a previsão do EHC de taludes em aterro foi desenvolvido através da aplicação de ferramentas de DM. Atualmente existe disponíveis diversos algoritmos de DM que podem ser aplicados a problemas de regressão ou de classificação. Neste estudo foram treinados dois algoritmos de referência na abordagem de problemas de regressão e classificação, nomeadamente as RNAs e as MVSs. Ambos os algoritmos têm evidenciado elevada eficiência na resolução de problemas reais (Tinoco et al., 2011), (Tinoco et al., 2014b), (Freitas et al., 2015).

As RNAs procuram imitar alguns aspetos do cérebro humano (Kenig et al., 2001), como o processamento de informação através da iteração entre vários neurónios (Perzyk e Kochanski, 2001). Neste trabalho adotou-se o modelo de RNA do tipo unidirecional e multicamada (*multilayer perceptron*), com uma camada intermédia e com *H* unidades de processamento. Controlando o valor de *H* podem ser realizadas análises mais complexas, ainda que um valor elevado de *H* poderá levar a um sobre-ajustamento do modelo aos dados de treino e consequente perda de capacidade de generalização deste. Para ultrapassar esta questão, o valor de *H* foi definido utilizando uma procura em grelha $\{0,2,4,6,8\}$ (H = 0 corresponde a uma regressão múltipla).

As MVS (Cortes e Vapnik, 1995), inicialmente desenvolvidas para problemas de classificação, foram mais tarde também aplicadas a problemas de regressão (Smola e Scholkopf, 2004) após a introdução da função perda ε-insensitiva. As MVS apresentam vantagens teóricas sobre as RNA tais como a ausência de mínimos locais durante a fase de aprendizagem. Isto é, estes modelos convergem sempre para a solução ótima. A ideia subjacente a uma MVS é transformar os dados de entrada num espaço característico de elevada dimensão usando um mapeamento não linear. Posteriormente, a MVS encontra o melhor hiperplano dentro do espaço característico. Esta transformação depende da função *kernel* adotada. O *kernel* Gaussiano é o mais popular por apresentar um menor número de parâmetros, tendo sido adaptado no presente estudo. Para auxiliar a escolha dos valores dos diferentes hiperparâmetros (γ, C e ε), foram adotadas as heurísticas propostas por Cherkassky e Ma (2004). Assim, para C foi adotado o valor de C=3 e a largura da zona ε-insensitiva foi definida de acordo com $\varepsilon = \hat{\sigma}/\sqrt{N}$, onde $\hat{\sigma} = 1.5/N \times \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$, $\hat{y}_i$ é o valor previsto pelo algoritmo dos 3-vizinhos próximos e N representa o número de registos da base de dados. O parâmetro *kernel* γ foi definido usando uma procura em grelha entre $\{1,3,5,7,9\}$.

Como ilustrado anteriormente na Figura 1, a distribuição dos registos pelas 4 classes EHC apresenta uma forte assimetria, aspeto que tem um efeito preponderante no desempenho dos algoritmos de aprendizagem. No sentido de ultrapassar esta questão, foram aplicadas duas abordagens visando o balanceamento da base de dados antes de iniciar o processo de aprendizagem. Assim, foram aplicadas as abordagens SMOTE (*Synthetic Minority Over-sampling Technique*) e sobre-amostragem (*Oversampling*) O SMOTE (Chawla et al., 2002) permite criar uma "nova" base de dados através da criação de novos registos tendo por base registos semelhantes (k vizinhos próximos). Esta estratégia é aplicada à classe minoritária. Simultaneamente são também removidos alguns registos da classe maioritária. Embora o SMOTE seja uma abordagem direcionada a problemas de classificação, Torgo et al., (2015) adaptaram esta metodologia a problemas de regressão. A sobre-amostragem corresponde a uma simplificação do SMOTE onde, aleatoriamente, registos da classe minoritária são repetidos para que todas as classes fiquem com o mesmo número de registos.

A avaliação do desempenho dos modelos foi realizada através do cálculo de diferentes métricas, nomeadamente (Baía, 2015): pontuação média de utilidade (PMU); precisão e exatidão. A PMU permite bonificar ou penalizar uma determinada classe em detrimento de outra. Assim, para o cálculo da PMU considerou-se a seguinte matriz custo-benefício:

Quadro 1- Matriz custo-benefício

Observado/Previsto	A	B	C	D
A	1	-4	-8	-16
B	-2	1	-4	-8
C	-4	-2	1	-4
DE	-8	-4	-2	1

A ideia subjacente à matriz apresentada no Quadro 1 consiste em penalizar qualquer classificação não correta, distinguindo se o erro é por excesso ou defeito. Por exemplo, prever um registo como D quando ele é A (penalização de -8) é menos penalizado em comparação a uma previsão de A quando o real nível de estabilidade é D (penalização de -16). Para todas as métricas quanto maior o valor da métrica melhor o desempenho do modelo. O PMU pode apresentar valores negativos (se em média as previsões representarem um custo) e o modelo ideal apresentará um PMU de 1. As restantes métricas, exatidão e precisão, podem variar entre 0% e 100%.

A capacidade de generalização dos modelos foi avaliada através da aplicação de uma validação cruzada com 5-fold (Hastie et al., 2009) e repetição de cada experiencia 20 vezes.

Todas as experiências foram conduzidas no ambiente estatístico R (R Development Core Team, 2009), com o auxílio do pacote rminer (Cortez, 2010), o qual é particularmente adequado para o treino dos algoritmos RNAs e MVSs.

4 - RESULTADOS

4.1 - Classificação nominal

Abordando a previsão do EHC como um problema de classificação, os algoritmos RNAs e MVSs foram treinados com a base de dados original mas também com uma base de dados balanceada (igual número de registos por classe) resultante da aplicação das abordagens SMOTE e sobre-amostragem.

O Quadro 2 compara o desempenho dos algoritmos RNAs e MVSs na previsão de EHC utilizando três métricas distintas. É também comparada a influência do balanceamento da base de dados através das abordagens SMOTE e sobre-amostragem.

Quadro 2 – Métricas de desempenho dos modelos – classificação nominal (melhores valores a negrito)

Modelo		PMU	Exatidão & Precisão			
			A	B	C	D
RNA	Normal	0.28	94.14 & 91.05	68.53 & **69.36**	66.22 & **68.55**	45.29 & 69.18
	SMOTed	0.18	86.31 & 93.88	72.75 & 59.34	48.85 & 63.03	**65.93** & 35.07
	OVERed	0.24	86.25 & 94.33	67.59 & 60.12	**67.12** & 57.87	**65.29** & 50.05
MVS	Normal	0.08	**95.03** & 88.59	64.89 & 64.66	49.76 & 62.41	0.55 & 77.17
	SMOTed	-0.12	76.39 & 94.17	72.23 & 49.13	47.26 & 45.13	37.94 & 33.22
	OVERed	-0.35	89.82 & 81.75	53.72 & 55.39	38.75 & 58.26	14.00 & 57.67

A Figura 2 ilustra e compara o desempenho dos algoritmos RNAs e MVSs, bem como o efeito do balanceamento da base de dados através das abordagens SMOTE e sobre-amostragem para os quatro modelos com a melhor resposta na previsão do EHC de taludes em aterro. Em cada um dos gráficos, as quatro barras representam as classes observadas e a graduação do preenchimento corresponde à classe prevista pelo modelo. Por exemplo, na Figura 2a (RNAs sem balanceamento da base de dados), mais de 68% dos registos da classe B foram corretamente previstos como pertencendo à classe B, menos de 25% foram classificados com a classe A e os restantes (cerca de 7%) como pertencendo à classe C (nenhum caso foi previsto como D).

Da análise conjunta do Quadro 2 e da Figura 2 observa-se um excelente desempenho na identificação dos taludes em aterro pertencentes à classe A (exatidão superior 90%), em particular nas situações sem balanceamento da base de dados. Para as restantes classes, embora se observe uma ligeira diminuição na resposta dos modelos, o desempenho obtido continua bastante elevado. De sublinhar os valores de exatidão superiores a 65% para as classes C e D, alcançados através do balanceamento da base de dados pela abordagem sobre-amostragem a aplicação do algoritmo RNAs. Comparando os algoritmos RNA e MVS, o primeiro apresenta um desempenho bastante superior, nomeadamente para as classes C e D, para as quais a probabilidade de rotura é superior.

Analisando o efeito das abordagens de balanceamento da base de dados SMOTE e sobre-amostragem, observa-se uma melhoria na resposta dos modelos na identificação dos taludes da classe D, em particular com a utilização do algoritmo RNAs. Relativamente às outras classes, o ganho é aproximadamente residual sendo mesmo em alguns casos negativo. Por exemplo, para a classe A e de acordo com o algoritmo RNAs observa-se uma diminuição da exatidão de 94% para 86% quando é aplicada uma abordagem de balanceamento da base de dados. Comparando o efeito global das duas abordagens de balanceamento da base de dados, a sobre-amostragem apresenta ser mais efetiva que o SMOTE.

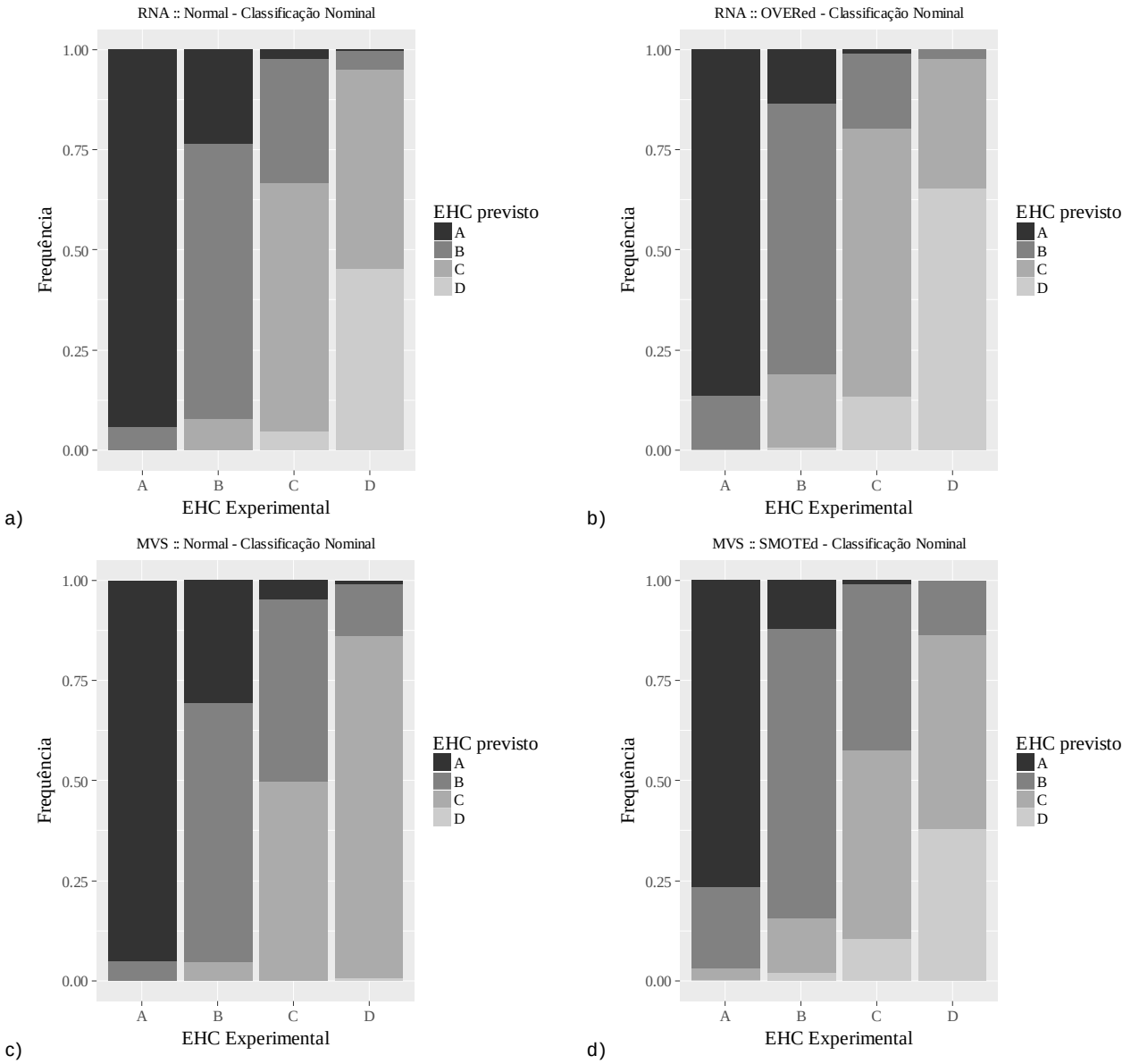


Figura 2 – Comparação do desempenho dos modelos – classificação nominal: a) RNAs sem balanceamento da base de dados; b) RNAs com sobre-amostragem; c) MVSs sem balanceamento da base de dados; b) MVSs com SMOTE;

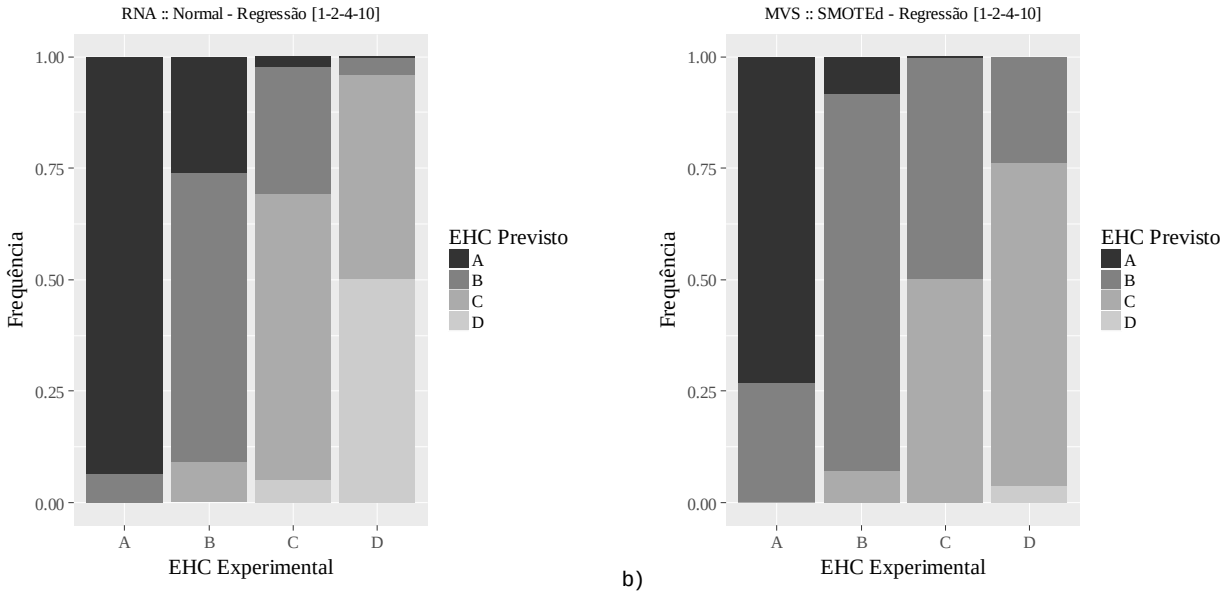
4.2 - Regressão

Tendo como objetivo melhorar o desempenho dos modelos, o problema em estudo foi convertido num problema de regressão e resolvido como tal. Para o efeito foram selecionadas diferentes escalas de regressão, tendo-se no final adotado a escala A=1, B=2, C=4 e D=10.

O Quadro 3 e a Figura 3 mostram e comparam o desempenho dos modelos de regressão na previsão do EHC de taludes em aterro. Também aqui as RNAs apresentam um desempenho superior às MVSs, em particular para as classes C e D. Por outro lado, e à semelhança dos modelos de classificação nominal, observa-se uma elevada exatidão na identificação de talude em aterro da classe A, com valores da exatidão superiores a 90%. Para as restantes classes o desempenho dos modelos diminui ligeiramente, observando-se ainda assim valores da exatidão superiores a 64% para as classes B e C e perto de 50% para a classe D, de acordo com o algoritmo RNA. Em particular para a classe D, a estratégia de classificação nominal mostrou ser mais eficiente na previsão do EHC de taludes em aterro. A Figura 3 ilustra a relação entre os valores observados e previstos de EHC de acordo com os melhores modelos RNA e MVS, evidenciando a melhor resposta das RNAs. A maior diferença de desempenho é observada para a classe D onde as MVS apresentam alguma dificuldade em identificar corretamente os taludes em aterro pertencentes a esta classe. O balanceamento da base de dados através da aplicação das abordagens sobre-amostragem e SMOTE apresenta um efeito pouco expressivo mesmo para as classes minoritárias.

Quadro 3 – Métricas de desempenho dos modelos – regressão

Modelo		PMU	Exatidão e Precisão			
			A	B	C	D
RNA	Normal	0.43	93.53 & 90.23	64.53 & 67.89	64.38 & 67.27	50.33 & 69.30
	SMOTed	0.44	90.21 & 92.60	71.00 & 64.40	**67.91** & 65.37	40.43 & 77.92
MVS	Normal	0.45	86.40 & 93.58	82.60 & 55.34	36.94 & 60.79	0.08 & **100**
	SMOTed	0.35	73.01 & **95.91**	**84.59** & 46.55	50.21 & 59.86	3.71 & 89.66



a) b)

Figura 3 – Comparação do desempenho dos modelos – regressão; a) RNAs sem balanceamento da base de dados; b) MVSs com SMOTE;

5 - OBSERVAÇÕES FINAIS

Neste trabalho foi apresentada uma proposta para a previsão do nível de estabilidade de taludes em aterro (EHC *Earthwork Hazard Category*), avaliado por quatro classes (A, B, C e D), através da aplicação de ferramentas de data mining e considerando como dados de entrada do modelo informação usualmente recolhida durante as inspeções de rotina (informação visual). Os resultados obtidos demonstram um desempenho muito promissor, tendo-se conseguido valores de exatidão superiores a 91% para classe A, aproximadamente 67% para as restantes classes. Verificou-se também que as Redes Neuronais Artificiais (RNA) apresentam um melhor desempenho na previsão do EHC comparativamente às Maquinas de Vetores de Suporte (MVS). Por outro lado, a aplicação de abordagens de balanceamento da base de dados, em particular a sobre-amostragem, permite uma melhoria do desempenho dos modelos, nomeadamente na previsão dos taludes da classe minoritária D. Observou-se ainda que abordando a previsão do EHC de taludes em aterro como um problema de classificação nominal é ligeiramente mais eficiente que seguindo uma abordagem de regressão.

Em jeito de observação final, gostaríamos de referir que o desempenho global obtido na previsão do EHC de taludes em aterro abre boas expectativas ao desenvolvimento de trabalhos futuros. Em particular, e tendo em conta o elevado número de variáveis utilizadas como atributos dos modelos, em trabalhos futuros pretende-se reduzir o número variáveis consideradas através da aplicação de metodologias de seleção de variáveis (e.g., recorrendo a técnicas de otimização como os algoritmos genéticos). Com isto pretender-se-á reduzir a complexidade dos modelos e eventualmente melhorar o desempenho dos mesmos. Será também de equacionar em desenvolvimentos futuros a aplicação de uma aprendizagem não supervisionada. Desta forma pretende-se minimizar a subjetividade inerente aos modelos apresentados, no que diz respeito à influência da equipa de Engenheiros e Técnicos especialistas na classificação atribuída a cada talude, a qual poderá não ser representativa do real nível de estabilidade do mesmo.

AGRADECIMENTOS

Este trabalho foi financiado pela FCT - "Fundação para a Ciência e a Tecnologia", no âmbito do ISISE, projeto: UID/ECI/04029/2013 e no âmbito do projeto: UID/CEC/00319/2013, bem como através da bolsa de pós-doutoramento com a referência SFRH/BPD/94792/2013 (POCH e FSE). Este trabalho foi também

financiado pelo COMPETE: POCI-01-0145-FEDER-007043. Um agradecimento especial ao Professor David Toll pela colaboração no desenvolvimento dos trabalhos, bem como à NetworkRail pela disponibilização da informação utilizada no presente estudo.

REFERÊNCIAS

AGC (2007). A national landslide risk management framework for Australia, *Australian Geomechanics Society*, Vol. 42, No. 1, pp. 1–12.

Ahangar-Asr, A., Faramarzi, A. e Javadi, A. (2010). A new approach for prediction of the stability of soil and rock slopes, Engineering Computations, Vol. 27, No. 7, pp. 878 sd–893.

Baía, L. (2015). *Actionable forecasting and activity monitoring: applications to financial trading*, Master's thesis, Faculdade de Ciências, Universidade do Porto, Porto, Portugal.

Chawla, N., Bowyer, K., Hall, L. e Kegelmeyer, W. (2002). Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research, pp. 321–357.

Cheng, M. e Hoang, N. (2014). Slope collapse prediction using Bayesian framework with k-nearest neighbor density estimation: Case study in Taiwan, Journal of Computing in Civil Engineering, pp. 1–8.

Cheng, M., Roy, A. e Chen, K. (2012). Evolutionary risk preference inference model using fuzzy support vector machine for road slope collapse prediction, Expert Systems with Applications, Vol. 39, No. 2, pp. 1737–1746.

Cherkassky, V. e Ma, Y. (2004). Practical Selection of SVM Parameters and Noise Estimation for SVM Regression, Neural Networks, Vol. 17, No. 1, pp. 113–126.

Cortes, C., e Vapnik, V. (1995). *Support Vector Networks, Machine Learning*, Vol. 20, No. 3, pp. 273–297.

Cortez, P. (2010). Data Mining with Neural Networks and Support Vector Machines using the R/rminer Tool, In P. Perner (Ed.), *Advances in Data Mining - Applications and Theoretical Aspects, 10th Industrial Conference on Data Mining*, LNAI 6171, Springer, pp. 572-583, Berlin, Germany.

Fay, L., Akin, M. e Shi, X. (2012). Cost-efective and Sustainable Road Slope Stabilization and Erosion Control, *Transportation Research Board*, vol. 430.

Freitas,E., Tinoco,J., Soares, F., Costa, J., Cortez, P., e Perreira, P. (2015). Modelling tyre-road noise with data mining techniques, Archives of Acoustics, Vol. 40, No. 4, pp. 547–560.

Gavin, K. e Xue, J. (2009). Use of a genetic algorithm to perform reliability analysis of unsaturated soil slopes, Geotechnique, Vol. 59, No. 6, pp. 545–549.

Hastie, T., Tibshirani, R. e Friedman, J. (2009). T*he Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer-Verlag, 745 p.

Kenig, S., Ben-David, A., Orner, M., e Sadeh, A. (2001). Control of properties in injection molding by neural networks, Engineering Applications of Artificial Intelligence, Vol. 14, pp. 819–823.

Lu, P. e Rosenbaum, M. (2003). Artificial neural networks and grey systems for the prediction of slope stability, Natural Hazards, Vol. 30, No. 3, pp. 383–398.

Miranda, T., Correia, A.G., Santos, M., Sousa, L.R. e Cortez, P. (2011). New models for strength and deformability parameter calculation in rock masses using data-mining techniques, International Journal of Geomechanics, Vol. 11, pp. 44–58.

Perzyk, M., e Kochanski, A. (2001). Prediction of ductile cast iron quality by artificial neural Networks, Journal of Materials Processing Technology. Vol. 109, pp. 305–307.

Pinheiro, M., Sanches, S., Miranda, T., Neves, A., Tinoco, J., Ferreira, A. e Gomes Correia, A. (2015). A new empirical system for rock slope stability analysis in exploitation stage, International Journal of Rock Mechanics and Mining Sciences, Vol. 76, pp. 182–191.

R DEVELOPMENT CORE TEAM (2009). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, Web site: http://www.r-project.org/, acedido em 28/02/2016.

Sakellariou, M. e Ferentinou, m. (2005). A study of slope stability prediction using neural networks, Geotechnical & Geological Engineering, Vol. 23, No. 4, pp. 419–445.

Smola, A. e Scholkopf, B. (2004). A tutorial on support vector regression, Statistics and Computing, Vol. 14, pp. 199–222.

Tinoco, J., Gomes Correia, A. e Cortez, P. (2011). Application of data mining techniques in the estimation of the uniaxial compressive strength of jet grouting columns over time, Construction and Building Materials, Vol. 25,No. 3, pp. 1257–1262.

Tinoco, J., Gomes Correia, A. e Cortez, P. (2014a). A novel approach to predicting young's modulus of jet grouting laboratory formulations over time using data mining techniques, Engineering Geology, Vol. 169, pp. 50–60.

Tinoco, J., Gomes Correia, A. e Cortez, P. (2014b). Support vector machines applied to uniaxial compressive strength prediction of jet grouting columns, Computers and Geotechnics, Vol. 55, pp. 132–140.

Torgo, L., Branco, P., Ribeiro, R. e Pfahringer, B. (2015). Resampling strategies for regression, Expert Systems, Vol. 32, No. 3, pp. 465–476.

Wang, H., Xu, W. e Xu, R. (2005). Slope stability evaluation using back propagation neural networks, Engineering Geology, Vol. 80, No3, pp. 302-315.

Yao,X., Tham, L. e Dai, F. (2008). Landslide susceptibility mapping based on support vector machine: a case study on natural slopes of hong kong, China, Geomorphology, Vol. 101, No. 4, pp. 572–582.